CUSLA - An Adaptive Machine Learning Algorithm Approximating the Subjective Quality of Chatbot Interaction.

Adrian Kenneally Software Design & AI Athlone Institute of Technology Athlone Ireland	Lukasz Jaczewski Software Design & AI Athlone Institute of Technology Athlone Ireland	Oisin Carley Software Design & AI Athlone Institute of Technology Athlone Ireland

ABSTRACT

This document presents CUSLA – An adaptive chatbot machine learning algorithm approximating the subjective quality of chatbot interaction. The CUSLA algorithm coordinates chatbots in adaptively altering their personalities when user satisfaction is low. Three chatbots were developed, each with a distinct language file of varying verbosity. A dataset of 900 interactions with the chatbots was recorded, including a user satisfaction rating for each. Our experiments show that user satisfaction with grows with the increasing verbosity.

KEYWORDS

Machine Learning, Subjective Quality Estimation, Chatbot, Emphatic Computing.

1. INTRODUCTION

University websites now hold more information than ever before. The move toward self-administration of services by students and staff alike, coupled with the advent of remote learning has seen the university website become an integral part of what is, or once was, campus life. The addition of chatbots to university websites is becoming more common. These chatbots come in many varieties, ranging from those that direct a user to an endpoint [1], answer FAQ's [2], or deliver a personalised experience based on user's behaviour [3].

The Principles of Universal Design (PUD) offer guidance on the design of products, environments, and communications [4]. Principle 1 suggests applications should provide the same experience to all users while Principle 2 allows for flexibility of use. Our chosen chatbot satisfies the former, providing the same experience to all.

This project demonstrates An Adaptive Machine Learning Algorithm Approximating the Subjective Quality of Chatbot Interaction. (CUSLA) which can be used to coordinate adaptive chatbots to change their personality type, language files, or other defining features. CUSLA works by estimating a user's current satisfaction rating of interaction with a chatbot. Once CUSLA's 1 to 5 rating falls below a threshold value a chatbot adaption is required. For the purpose of this project the adaption focuses on the perceived "friendliness" of the chatbot responses.

Three chatbots were developed each with a language file containing responses of different verbosity. The least verbose file is marked as “direct” and contains mostly one word answers. The “normal” language file contains more pleasantries, with responses of a type that would be commonly expected from a chatbot. The final file is highly verbose returning the user more text while attempting to present itself in a “friendly” manner. Through a brute force test the user preference for the “friendly” language file was established.

II details current chatbot technology and explores current adaptive chatbots and how they can be applied to chatbot design. Methods of sentiment analyses, another component of the CUSLA are also explored. III gives an overview of CUSLA and the chatbot architecture, V discusses results while conclusions and the direction of further work is considered in VI.

2. LITERATURE REVIEW

2.1 Existing Chatbot Technology

In [1] the authors chatbot, DEXTER, converses with users, providing suitable answers to queries. When presented with a question, a database check for a suitable response is made. If found, the suitable response is relayed to the user. When not found, a response is generated through pattern matching and AI. A notification alerting admin to add this query to the database is also generated. Dexter is built on the open source RASA framework. RASA provides natural language understanding, and dialogue management through integration with Long Short Term Memory (LSTM,) a form of Recurrent Neural Network (RNN.)

In [5] the authors demonstrate a FAQ’s answering system. Here, the model was trained on a corpus of conversations from movies, with the Bag-of-words technique, applied to find the frequency of a word in a given sentence. Then, by adding a seq2seq model for converting sequences from one domain to another, the chatbot was able to extract features from a given FAQ text.

[6] details the design of a virtual assistant “Wisdom,” built on the FLASK Framework. FLASK acts as an intermediary between the UI and a Natural Language Processing (NLP) module. The NLP module pretrained on word2vec, again uses the bag-of-words technique, with spaCy, and various TensorFlow to generate intent and entities from user input. Depending on the sentence type, a web scraping or ML function is called. Similar to [5] LSTM helps produce an output.

2.2 Adaptive Chatbots

PUD [5] offer guidance on the design of applications, environments, and communications. A set of architectural guidelines, they provide valid considerations when designing a chatbot. Principle 1 centres on “Equitable Use” and when applied to chatbot design indicates that interaction should be identical for all users. However, Principle 2 sets out the expectation that an application be flexible and “Provide adaptability to the user’s pace.” This encourages personalization of a chatbot. It could be expected that a chatbot adhering to Principle 2 would be capable of adapting to the user’s requirements.

The balance between the two Principles must be given consideration. [9] presents a case in favour of adhering to Principle 1, stating interactions with more human like chatbots can increase users’ expectations, leading to frustration with chatbot failures. A more generic, task oriented chatbot would not incite such frustrations. It is our belief that the addition of personality to a chatbot infringes Principle 1 of PUD. However, article [7] highlights increased user engagement with personalised chatbots. In addition, users subconsciously assign personality to chatbots during interaction [8]. Study of the effect of a chatbots perceived personality is carried out in paper [14].

This study’s goal was to establish if users have a preference for a specific chatbot personality type. Two text based chatbots were developed to use separate language files specifically selected to convey personality types from the Five Factor Model. By basing the chatbots purely on text interactions, the paper’s authors reduced the possibility of persona related effects influencing user's perception of the chatbot's personality. No voice or visual representations of the chatbot were included. Chatbot A was designed to exhibit high extraversion and agreeableness. Its responses contained high energy punctuation, had a highly verbose nature, and commonly uses complementary language. Chatbot B was designed in so far as possible to be the opposite in each. Before interaction with the chatbot, testing participants completed a Big-Five Inventory type survey, identifying each participant's personality traits. They were then asked to interact with both chatbots. Users' opinions of the chatbots were then established through further questionnaires. This study found that, although participants conversed more with Chatbot B which exhibited less agreeableness and extraversion, a higher proportion of users preferred Chatbot A. Also of note was that no strong correlation was found between agent preference and the participants predominant personality traits.

2.3 The need for Sentiment Analyses

Section 2.1 suggests, albeit while carrying some considerations, that the design of a chatbot may be improved if that chatbot were to adapt its personality to the end user's personality. Identifying the end user's personality is a non-trivial task. Matching personalities to users is not the only scenario in which it may be appropriate for a chatbot to display adaptive features. Current user sentiment may also be considered, what if a user interacting with a chatbot personality type was displaying frustration or anger? Could this be used to induce a change in the chatbot?

Article [10] presents a chatbot which interestingly, contains a sentiment analyses tool to estimate the tone and current sentiment of user inputs. The UMLFit classifier was used to build a fine grain sentiment analyses module capable of classifying thirteen emotions including joy, anger, love, and sorrow among others. A suggested use case of the sentiment analyses module, directing students whose inputs have been classed as sorrowful toward counselling services could add great value to university chatbot project. If such developments are deemed out of scope of this project the previous iteration of this chatbot, presented in article [11] provides an alternative possibility. Here a sentiment analyses of the chatbot output returns a positive neutral or negative rating, based on that score appropriate animations and emojis are activated on the user interface. This methodology could be used in an algorithm current sentiment of a user.

Further to the notion of a positive-neutral-negative user input sentiment article [12] utilises VADER, a lexicon and rule-based sentiment analyses tool that is specifically tuned to sentiments expressed in social media posts, a domain with user statements similar to what may be expected to our input. VADER was chosen for its pace of processing and lightweight nature not requiring training data.

3. IMPLEMENTATION

3.1 Chatbot architecture

The chatbot has been preconfigured responses related to the AIT website mapped to certain patterns. The Natural Language Processing Module is responsible for user utterance and intent classification. Users' text inputs are stemmed and tokenized into structured data. The machine learning module's feed forward neural network, containing two layers, is provided a numerical input from the bag-of-words. Output from the neural network is the class (tag from JSON file) to which the bag-of-words belong. Based on this class a corresponding response is returned to the user.

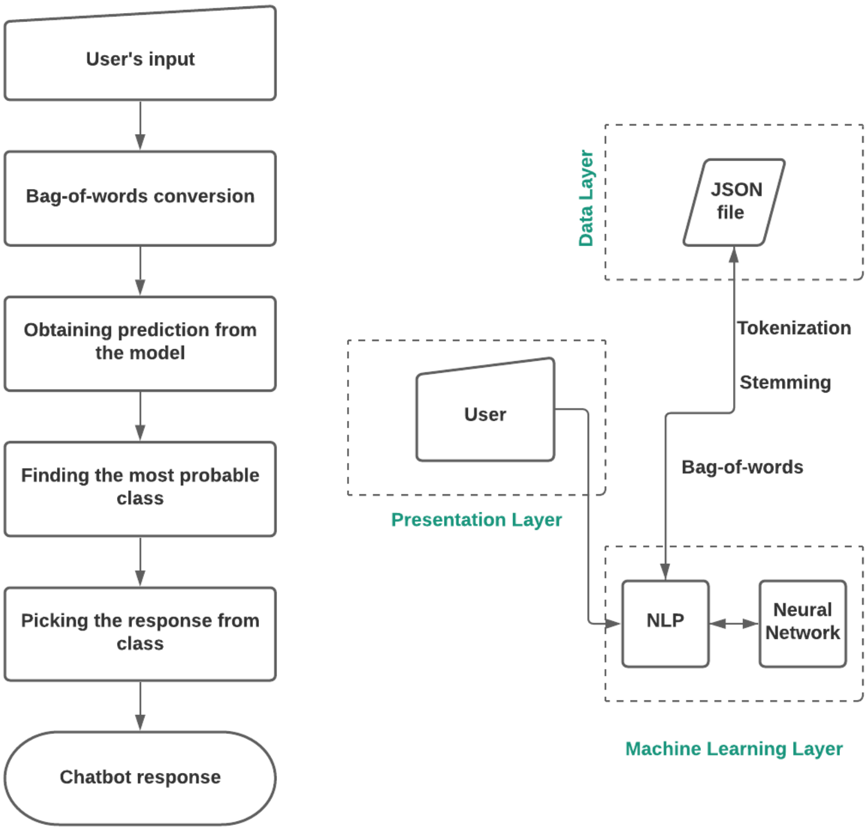


Figure 1 **Chatbot Architecture**

3.2 CUSLA algorithm

A Multi-layer perceptron neural network is used as the classifying algorithm. It has 100 hidden layers. This supervised machine learning algorithm learns from 720 dataset records and is then tested on the remaining 180. It outputs a discrete value in this case "star rating" based on the other input metrics. There are five separate clusters to which the algorithm can output to, being the different star ratings. A Multi-layer perceptron was used because of how well they outperform other machine learning algorithms on both classification and regression tasks when they are tuned and trained properly.

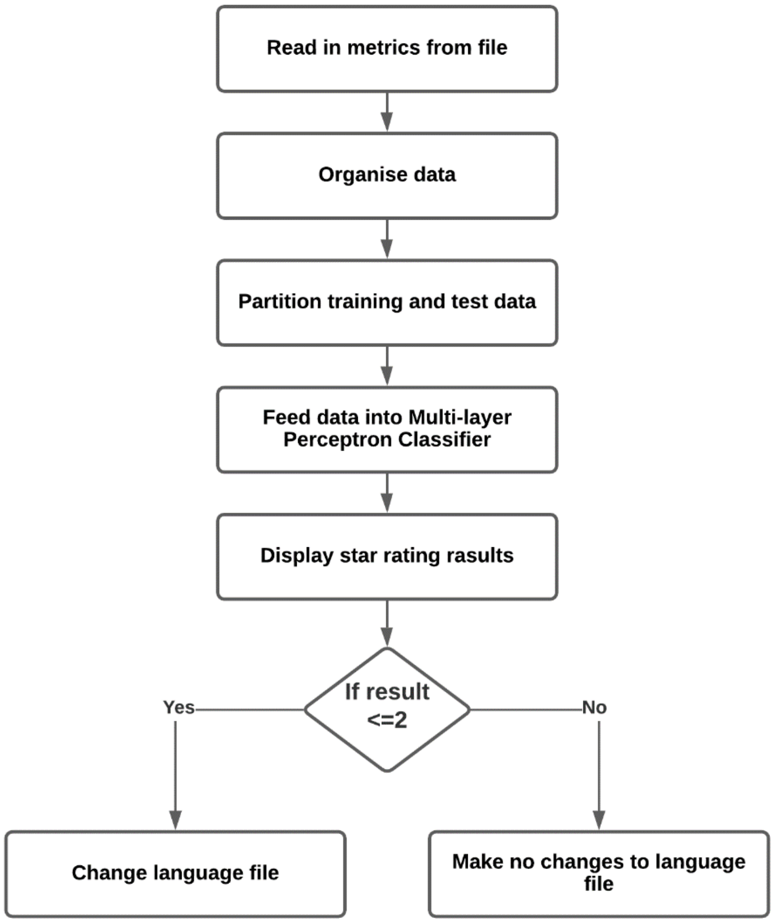

Read in metrics from file
Organise data
Partition training and test data
Feed data into Multi-layer Perceptron Classifier
Display star rating rasults
If result <=2
Yes
No
Change language file
Make no changes to language file

Figure 2 CUSLA Flowchart

CUSLA Pseudocode:

```
# input section
        data = load the data from data_chat.csv file

# organise data section
        df = create DataFrame from data
        X = create an array from df[“messages”, “time”, “sentiment”, “goodbye”]
        X = reflect the X over its main diagonal by writing rows as columns and vice-versa using
        transpose()
        Y = create an array from df[“stars”] (containing column ‘stars’ only)
        X_train, X_test, Y_train, Y_test = split arrays X, Y into train and test subset

# processing section
        mlp = implement Multi-layer Perceptron classifier
        Y_pred = predict class labels for the provided X_test data
        Y_score = return the mean accuracy on the given test data and labels
        result = extract the predicted star rating from Y_score

# output section
        display “Test set predictions:” + Y_pred
        display "Test set accuracy:" + Y_score

# result section
        if result <= 2
                change_language_file()
```

Table 1 CUSLA Pseudocode

4. RESULTS/DISCUSSION OF CLASSIFICATION OF USER EXPERIENCE

5.1 Brute Force Test

In order to develop a baseline satisfaction score per chatbot type a brute force test of each was executed. Three hundred sample interactions were recorded and scored per chatbot. Users, familiar with the AIT website, were tasked with reaching an endpoint of their choosing via the chatbot. On conclusion, the user was asked to submit a one-to-five-star satisfaction rating. Time outliers greater than two standard deviations from the mean were removed. Table 3 displays the average metrics achieved per chatbot type. Table 2 displays the sample conversation with NLP based chatbot conducted during our research. The user who interacted with chatbot rated this conversation 4 out of 5.

Start talking with the bot (type quit to stop)!

Student: Hi. How are you?

Chatbot: Hi there, how can I help?

Student: What faculties do you have at your university?

Chatbot: Current Faculties; Business & Hospitality, Engineering & Informatics, Science & Health, Continuing Professional & Online and Distance Learning

Student: Thanks.

Chatbot: Happy I could be of help.

Student: What kind of student clubs and societies do you have?

Chatbot: At Athlone Institute of Technology, we have 28 clubs and 23 societies, if you would like more information, please follow the link: https://www.ait.ie/life-at-ait/clubs-and-societies/

Student: What are the tuition fees?

Chatbot: Information about fees can be found here: https://www.ait.ie/life-at-ait/registry/fees

Student: I really appreciate it. It was very helpful.

Chatbot: Anything for you!

Student: Goodbye

Chatbot: Talk to you later.

Table 2 Sample Conversation with NLP based chatbot

5.1.1 Metrics Collected

Messages – the number of messages sent by the user. Time – the time the chatbot is active before the conversation is ended in seconds. URL Usefulness – proportion of URL responses the user found useful. Goodbye Sentiment – if the user exited the interaction in a courteous manner. Input Message Sentiment – the average sentiment of user input messages per chatbot interaction, input messages were given a sentiment score – 1 for negative, 0 for neutral and +1 for positive. Star Rating – rating the user gave the chatbot ranging from 1-5.

In Table 3 we see the normal chatbot had the most messages received and on average the longest time per user interaction. It is also tended to provide the correct URL most often between the 3 chatbots. The normal chat also had the highest average sentiment with .1 but the friendly chatbot had the highest star rating with 2.8.

	Direct	Normal	Friendly
Average Messages	7.85	8.093645	7.636667
Average Time	78.74	84.99	75.17
Average URL Usefulness	0.33	0.46	0.419
Average Input message sentiment	0.043	0.10	0.09
Average Star Rating	2.57	2.77	2.80

Table 3 Average of metrics collected for each chatbot

We can see that the most ratings received were for four-star ratings followed closely by 3- star ratings, we then can see that there was not a lot of 5-star ratings for chatbot which will lead to difficulty with training and testing the algorithm.

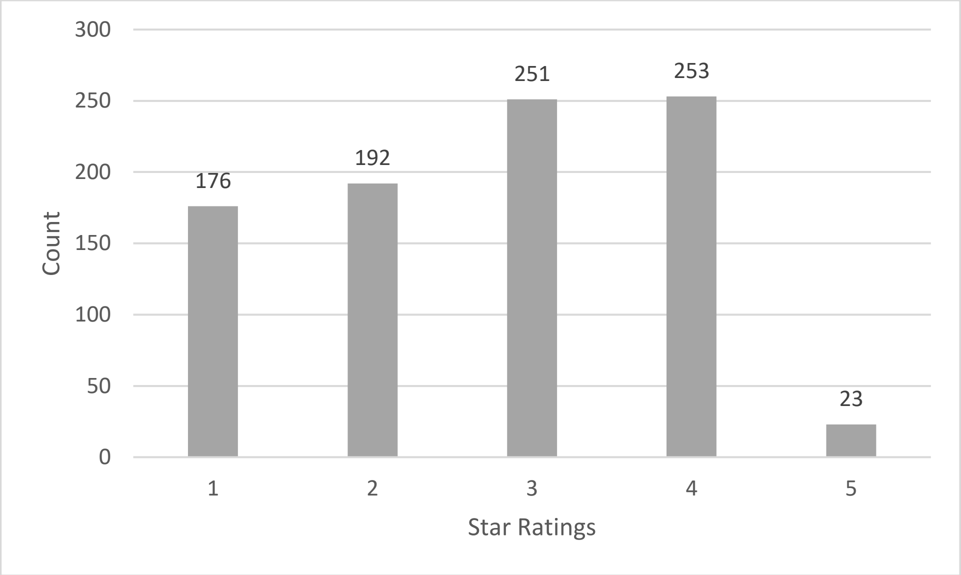


Figure 2 Count Star Ratings

Looking at figure 4 we can see that the direct chatbot had majority 4-star ratings (38%) followed closely by 3-star ratings (30%) and then 2-star ratings (20%). In both figure 5 and 6, 4-star rating is the highest but with the normal chatbot with have a higher percentage of 4-star ratings this lends itself to the higher average star rating for the normal chatbot.

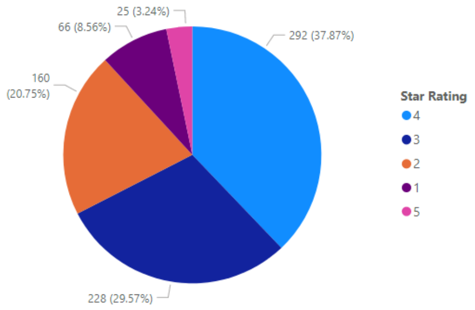


Figure 3 Direct Chatbot Star Ratings

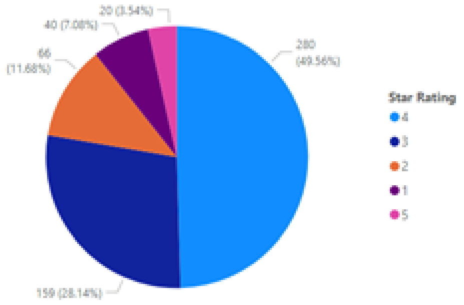


Figure 5 Normal Chatbot

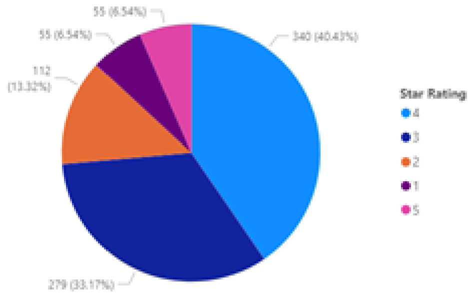


Figure 6 Friendly Chatbot

5.1.2 Brute Test Results

The data shows that most star ratings are 3 and 4. With very little 5-star ratings and comparatively little 1-star ratings this will affect the overall training of the algorithm. The friendly chatbot seems to have the highest user rating only by a fraction and the direct chatbot seems to have the lowest rating. This could lend to the idea that a chatbot must have some personality for the user to interact with the chatbot as the direct chatbot has the lowest user satisfaction.

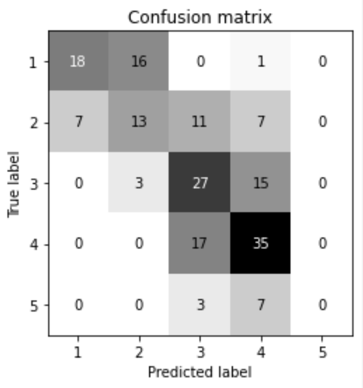
Again, the input message sentiment was highest with the normal chatbot but very close to the friendly chatbot. The direct chatbot has the lowest sentiment this could be because the user interacting with the chatbot might change their approach to a more direct one based on the interaction. The URL Usefulness had the highest score for the normal chatbot followed closely by the friendly chatbot and the direct. Each chatbot had a similar number of messages received and a similar interaction time. An interesting observation is that while the normal chatbot has the highest URL usefulness and input message sentiment, which could be interpreted as a better interaction, the friendly chatbot has the highest user satisfaction score. This seems to suggest even though the normal chatbot performs better and friendly chatbot leaves the user with a better "feeling" of the experience. It would be interesting to see if this trend would continue with more data.

5.2 CUSLA Algorithm results

Fig 7 is confusion matrix of CUSLA. The large proportion of 3- & 4-star ratings mean the algorithm is likely to return a 3 & 4-star rating. No 5-star ratings were predicted. In the test set the accuracy of the model was still only 0.55 and, in the training set the accuracy of the model was 0.61.

Stars	Precision	Recall	F1-score	support
1	0.76	0.52	0.62	42
2	0.46	0.4	0.43	47
3	0.41	0.74	0.53	54
4	0.58	0.48	0.52	69
5	0	0	0	13

Table 4 Classification Report



Confusion matrix

True label \ Predicted label	1	2	3	4	5
1	18	16	0	1	0
2	7	13	11	7	0
3	0	3	27	15	0
4	0	0	17	35	0
5	0	0	3	7	0

Figure 7 Confusion matrix

As seen in the classification report 1- star has the highest precision score, meaning when a 1 star is predicted it is correct 76% of the time, and 5 star has the lowest precision score as it is never predicted. This is probably due to the lack of 5-star reviews for the chatbot. 3-star reviews also have quite a low precision score as it is misclassified quite often.

Then 3-star ratings had the highest recall meaning that it correctly identifies 74% of all 3-star ratings. On the other hand, 2-star ratings are only correctly identified 40% of the time.

1-star ratings have the highest F1 score so there is the most accurate rating on the model. Looking at support we can see that 5-star ratings are vastly less than any other classifications. In order to have a better model we would need to address this problem.

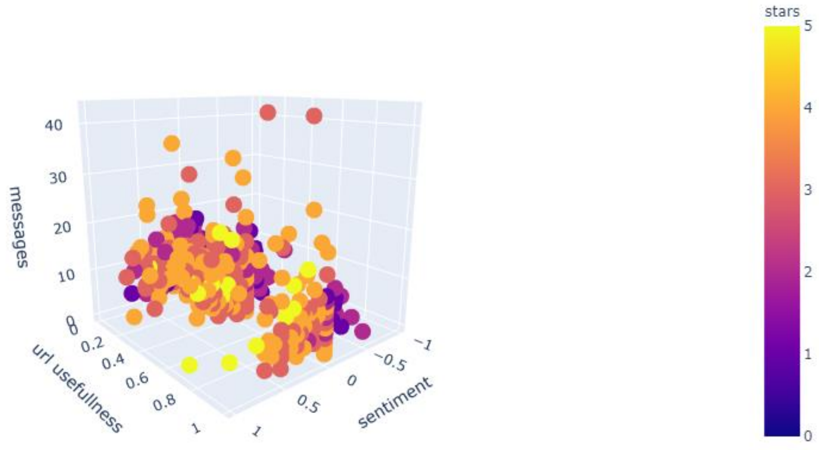


Figure 8 Comparison of Evaluation Metrics and star rating cluster

Fig 8 shows the rating a user gives can vary on several factors. Based on this graph one of the most dominant indicators of the user satisfaction score is “URL usefulness” if the user does not find the URL useful, they will often give the chatbot a rating of 1. A higher URL usefulness often results in a higher chatbot rating as seen from the graph with the darker colors of purple towards the back of the graph indicting a low user satisfaction and low URL usefulness. Another major influence on star rating is sentiment, again looking at the graph as you move from left to right you can see that the overall star rating increases with more lightly colored data points, showing a higher star rating.

5. CONCLUSION

The original aim of this paper was simply to create a chatbot through which users find information about the Athlone Institute of Technology website. The idea behind this was that due to the current pandemic the only access users could have to AIT was through the website and to help them access the information they needed. As we researched the literature, we found that there seemed to be contradictions in the way a chatbot should interact with users. With some authors claiming that a chatbot should mimic and user’s personality and reflect that based on attraction theory, while others theorized that the principles of universal design should be followed. This left an opportunity to explore if there is a best way to approach the matter of how a chatbot should interact.

The approach we decided to employ was that of an algorithm that could predict the user’s satisfaction with the interaction of the chatbot. After every interaction with the chatbot the algorithm could score the interaction. Based on this score the type of chatbot interaction could be changed, i.e friendly, direct or normal. While overall the user satisfactions ratings were relatively close, the friendly chatbot had the highest average user interaction. How the user scores their interaction with the chatbot is completely subjective, and changes depending on if the user would like just to get the answer to their question or if they would like to simply “chit-chat” with the chatbot. This should also be taken into consideration when observing the results.

Containing 900 overall interactions, the dataset is small to train an algorithm on. Future work to record more interactions will improve the overall quality and interaction of the chatbot. More data will also provide further opportunity to compare and contrast the interaction with different types of chatbots and if there is a benefit in a personalized or non-personalized chatbot.

This research provided a great base to move forward with more exploration of the topic of personalization of a chatbot. Obviously with more data we can get a clearer picture user satisfaction scores but there is also an opportunity to collect qualitative data on the users “feeling” of the experience. With this information it could greatly provide context and reasoning behind the scores collected. In the future in would be the hope of the group that the algorithm could be used to change the interaction of the chatbot in live time with the users interaction in order to provide the user with the greatest experience possible.

REFERENCES

[1] N. T. Thomas, "An e-business chatbot using AIML and LSA," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 2740-2742, doi: 10.1109/ICACCI.2016.7732476.

[2] A. Huddar, C. Bysani, C. Suchak, U. D. Kolekar and K. Upadhyaya, "Dexter the College FAQ Chatbot," 2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW), Mumbai, India, 2020, pp. 1-5, doi: 10.1109/ICCDW45521.2020.9318648.

[3] F. Grivokostopoulou, I. Perikos, M. Paraskevas and I. Hatzilygeroudis, "An Ontology-based Approach for User Modelling and Personalization in E-Learning Systems," 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 2019, pp. 1-6, doi: 10.1109/ICIS46139.2019.8940269.

[4] http://universaldesign.ie/What-is-Universal-Design/The-7-Principles/

[5] S. P. Reddy Karri and B. Santhosh Kumar, "Deep Learning Techniques for Implementation of Chatbots," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-5, doi: 10.1109/ICCCI48352.2020.9104143.

[6] S. V. Prajwal, G. Mamatha, P. Ravi, D. Manoj and S. K. Joisa, "Universal Semantic Web Assistant based on Sequence to Sequence Model and Natural Language Understanding," 2019 9th International Conference on Advances in Computing and Communication (ICACC), Kochi, India, 2019, pp. 110-115, doi: 10.1109/ICACC48162.2019.8986173.

[7] C. Steinmacher, A. Paula and G. Marco Aurelio, How should my chatbot interact? A survey on human-chatbot interaction design., 2019.

[8] A. Ho, J. Hancock and A. S. Miner, "Psychological, relational, and emotional effects of Self-Disclosure after conversations with a chatbot," Journal of Communication, 2018.

[9] T. Araujo, "Living up to the chatbot hype: The influence of anthropomorphic designcues and communicative agency framing on conversational agent and company perceptions.," Computers in Human Behaviour, 2018.

[10] S. R. Murali, S. Rangreji, S. Vinay and G. Srinivasa, "Automated NER, Sentiment Analysis and Toxic Comment Classification for a Goal-Oriented Chatbot," 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), Fez, Morocco, 2020, pp. 1-7, doi: 10.1109/ICDS50568.2020.9268680.

[11] H. V. Kumar et al., "PESUBot: An Empathetic Goal Oriented Chatbot," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 2018, pp. 1083-1089, doi: 10.1109/ICACCI.2018.8554916.

[12] C. W. Park and D. R. Seo, "Sentiment analysis of Twitter corpus related to artificial intelligence assistants," 2018 5th International Conference on Industrial Engineering and Applications (ICIEA), Singapore, 2018, pp. 495-498, doi: 10.1109/IEA.2018.8387151.

[13] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text", Proc. The Eighth International AAAI Conference on Weblogs and Social media, pp. 2-10, May. 2014.

[14] Ruane E., Farrell S., Ventresque A. (2021) User Perception of Text-Based Chatbot Personality. In: Følstad A. et al. (eds) Chatbot Research and Design. CONVERSATIONS 2020. Lecture Notes in Computer Science, vol 12604. Springer, Cham. https://doi.org/10.1007/978-3-030-68288-0_3